DOCUMENT RESUME

ED 442 839 TM 031 259

AUTHOR Monahan, Patrick

TITLE The Effect of Unequal Variances in the Ability Distributions on the Type I Error Rate of the Mantel-Haenszel Chi-Square Test for Detecting DIF.

PUB DATE 2000-04-00

NOTE 58p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 25-27, 2000).

PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC03 Plus Postage.

DESCRIPTORS *Ability; Chi Square; *Item Bias; Simulation; Statistical Distributions; *Test Items

IDENTIFIERS Item Bias Detection; *Mantel Haenszel Procedure; Type I Errors; *Variance (Statistical)

ABSTRACT

Previous studies that investigated the effect of unequal ability distributions on the Type I error (TIE) of the Mantel-Haenszel chi-square test for detecting differential item functioning (DIF) simulated ability distributions that differed only in means. This simulation study suggests that the magnitude of TIE inflation is increased, and the type of items that show inflation are somewhat broadened, under the realistic scenarios of ability distributions that differ not only in means but also in variances or variances alone. There were several conditions for which unequal variances in the ability distributions, either alone, or in combination with unequal means, produced practically important TWI inflation when no such inflation was observed under unequal means alone. This occurred primarily for the hard and medium difficulty items with high discrimination and on the short test with the total sample size of 2,000, but it was also observed for the medium-length test. The most peculiar finding was the aberrant behavior of the highly discriminating hard items, which was the only item to show marked TIE inflation due to unequal variances alone when only rare mild inflation was observed under the combination of unequal variances and unequal means. (Contains 14 tables, 5 figures, and 39 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 442 839

The Effect of Unequal Variances in the Ability Distributions on the Type I Error Rate of the Mantel-Haenszel Chi-square Test for Detecting DIF

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

P. Monahan

Patrick Monahan

The University of Iowa

Paper presented at the annual meeting of the National Council on Measurement in Education, April, 2000, New Orleans.

TM031259

Correspondence: Patrick Monahan, N476 Lindquist Center, University of Iowa, Iowa City, IA 52242. E-mail: pmonahan1@aol.com.

BEST COPY AVAILABLE

Abstract

Previous studies that investigated the effect of unequal ability distributions on the Type I error (TIE) of the Mantel-Haenszel chi-square test for detecting differential item functioning (DIF) simulated ability distributions that differed only in means. The present study suggests that the magnitude of TIE inflation is increased, and the type of items that show inflation are somewhat broadened, under the realistic scenarios of ability distributions that differ not only in means but also in variances, or variances alone. For example, the highly discriminating moderately difficult item (a=1.5, b=0) manifested TIE inflation under the combination of unequal means and unequal variances when no such inflation was observed under unequal means alone. Interestingly, the highly discriminating hard item (a=1.5, b=1) exhibited inflation under unequal variances alone; however, only rare minor inflation occurred under the combination of unequal variances and unequal means, and no inflation was noted under unequal means alone. When inflation was demonstrated under unequal means alone, as expected from previous research, for the easy item with high discrimination (a=1.5, b=-1), and the hard item with low discrimination (a=.5, b=1), the inflation was worse when the ability distributions also differed in variances. The inflation for these items was greatest on the short test (21 items versus 41 items), with larger total sample size (2000 versus 1000), and a Reference/Focal group sample size ratio of unity (1.0 versus 3.0). Previous studies had not systematically disentangled the effect of sample size ratio from that of total sample size. The sample size ratio of 1.0 produced more inflation than the 3.0 ratio, but only for the sample size of 2000. The relationship between TIE and area of overlap under the ability distributions was monotonic decreasing for all items except one (a=1.5, b=1). The relationship was remarkably linear for all items, but only within levels of the equality of ability means factor.

Introduction

The Mantel-Haenszel chi-square (χ^2_{MH}) test (Mantel & Haenszel, 1959) is one of the most commonly used tests for detecting differential item functioning (DIF). For *Row* x *Column* x *Strata* contingency tables, the generalized χ^2_{MH} (Birch, 1965; Landis, Heyman, & Koch, 1978; Mantel & Byar, 1978) has $(R\text{-}1)(C\text{–}1)$ degrees of freedom (df). In the context of DIF, it is a test of conditional independence between group (R) and item score (C), controlling for ability (S). Ability is usually represented by the total observed score. For 2 x 2 x S tables (df=1), the χ^2_{MH} is the uniformly most powerful test (Birch, 1964) of the null hypothesis that the common (or adjusted) population odds ratio (α) equals 1.0, against the specific focused alternative hypothesis that it is not equal to 1 *and constant* across the strata. The alternative hypothesis is that the odds of a correct response is different for the reference group (RG) than it is for the focal group (FG), controlling for total score. The Mantel-Haenszel delta-DIF (MH-D-DIF) is a log-transformation of the estimated Mantel-Haenszel common odds ratio [$-2.35\ln(\hat{\alpha}_{MH})$]. It is a descriptive measure of the magnitude of DIF and converts $\hat{\alpha}_{MH}$ to the logistic definition of the delta scale used at ETS to measure item difficulty (Holland & Thayer, 1985). Under the null hypothesis of no DIF, α_{MH} =1 and MH-D-DIF = 0.

Holland (1985) proposed using the χ^2_{MH} test for the analysis of DIF in dichotomous items between two groups (2 x 2 x S table). Holland and Thayer (1988) described the theory behind this approach. They and others (Shepard, Camilli, & Averill, 1981; Shepard, Camilli, & Williams, 1984) stated that IRT methods are theoretically preferred for analyzing DIF *if* the IRT model holds. Indeed, Lord (1977; 1980) provided a definition of DIF in terms of IRT that some believe is more theoretically fundamental than other measures of DIF (Donoghue, Holland, &

Thayer, 1993, p.140). Holland and Thayer (1988) showed theoretically that for the Rasch model, the null hypothesis, for which the χ^2_{MH} was developed, holds exactly in the population if the matching items exhibit no DIF (but the studied item may exhibit DIF), the matching score includes the studied item, and the data are random samples from the RG and FG. Under the Rasch model, if these three conditions hold, the null hypothesis for the IRT likelihood ratio test coincides with the null hypothesis for the χ^2_{MH} test. Others confirmed (Donoghue et al., 1993; Lewis, 1993; Merideth & Millsap, 1992) that, contrary to intuition, the studied item should be included in the total score. This is now an accepted practice.

Zwick (1990) proved theoretically that for non-Rasch models, unequal ability distributions produce inflated Type I error (TIE) even when the studied item is included. Simulation studies verified this fact as well as the occurrence of biased estimates of the Mantel-Haenszel adjusted odds ratio (Donoghue et al., 1993; Lu, 1996; Roussos & Stout, 1996; Uttaro & Millsap, 1994; Zwick, 1990). These studies revealed other factors contributing to inflated TIE: short test length, large sample size, characteristics of the studied item (such as high discrimination, low difficulty, and low pseudo-guessing), and low average core item discrimination. However, there were several interactions. As expected from theory (Zwick, 1990), equal means in the ability distributions yielded no inflated TIE for any of the above factors. Unequal means in the ability distributions produced inflation but only for certain items, especially easy items with high discrimination (e.g., a=1.5, b=-1), and to a lesser degree, hard items with low discrimination (e.g., a=.5, b=1), and primarily for short tests (e.g., 20 to 26 items). The TIE associated with these items increased when the easy items with high discrimination also had a low pseudo-guessing parameter (e.g., c=0) or when the hard items with low discrimination had a moderate pseudo-guessing parameter (e.g., c=.20). The TIE observed in

these studies also increased with larger sample size (Lu, 1996; Roussos & Stout, 1996), and lower average core item discrimination (Lu, 1996). The percentage of core items exhibiting DIF only had a small effect on TIE inflation (Donoghue et al., 1993; Narayanan & Swaminathan, 1994). The standard deviation of core item discrimination did not have an effect (Lu, 1996). Rogers and Swaminathan (1993) found no TIE inflation but they only simulated ability distributions that were identical, for which inflation is not expected. Narayanan and Swaminathan (1994) found small but practically unimportant inflation; however, they simulated medium length (40 items), not short tests, and the largest focal group sample size was 300. The parameters used by these simulation studies as well as the present study are shown (Table 1). In the studies above, the true abilities (theta values) of the RG examinees were randomly sampled from the $N(0,1)$ distribution, except for the study by Roussos and Stout where the values for μ_R and μ_F were chosen such that the midpoint between them equaled the average item difficulty.

However, in these previous studies, "unequal" ability distributions were simulated only as a difference in means. The ability distribution of the FG was simulated as either $N(-.5,1)$ or $N(-1,1)$, which is indeed of practical interest. However, for standardized achievement tests, the ability distributions for the RG and FG often differ not only in means but also in variances. For example, at the primary level, of 178 combinations of grade (3-8) and test, the variance ratio (VR) of *ITBS* (Hoover, Hieronymus, Frisbie, & Dunbar, 1993) scores for Whites to African-Americans ranged from .96 to 1.87, with a median of 1.38 (Table 2). At the secondary level, of 44 combinations of grade (9-12) and test, the same ratio of *ITED* (Feldt, Forsyth, Ansley, & Alnot, 1993) scores ranged from .91 to 1.63, with a median of 1.34 (Table 2). On the *ACT* college entrance exam (ACT, 1992), the ratio for the four tests ranged from 1.19 to 1.80. The

ratio of variances of scores for Males to Females ranged from .88 to 1.40 (median=1.11) on the *ITBS*, from .99 to 1.36 (median=1.21) on the *ITED*, and from 1.04 to 1.25 on the *ACT* (Table 2).

Bielinski and Davison (1998) noted several articles that brought attention to the importance of gender differences in variability, especially for mathematics scores (Benbow, 1988; Benbow & Stanley, 1980; Benbow & Stanley, 1983; Feingold, 1992; Feingold, 1994; Feingold, 1995; Hedges & Friedman, 1993; Hedges & Nowell, 1995; Humphreys, 1988). The authors then showed, using actual mathematics scores for representative eighth-grade students, that differences in variability can lead to a gender-by-item-difficulty interaction in which Males perform better than Females on the hardest items, and Females perform better than Males on the easiest items. The meta-analysis by Hedges and Nowell (1995) described the Male/Female variance ratios (VRs) of mental test scores from six large data sets that used national probability samples, collected between 1960 and 1992. The tests included reading comprehension, vocabulary, mathematics, perceptual speed, science, social studies, and nonverbal reasoning. Of the 37 combinations of data set and test, only two VRs were less than one (.82 and .98). The largest VRs (2.72 and 2.34) were for an electronics and an auto test, respectively. The remaining 33 VRs ranged between 1.00 and 1.74, and most were less than 1.28.

Therefore, the RG and FG often differ in the means or variances (or both) of their observed score distributions. Presumably, this holds as well for their ability distributions despite the inexact correspondence between these two types of distributions for non-Rasch models of item responses. The evidence suggests that the variance of achievement and cognitive ability test scores is usually greater for the RG than the FG. Nevertheless, there appears to be no published investigations of the effect of unequal variances in the ability distributions on the TIE rate of the χ^2_{MH} test. Pommerich, Spray, and Parshall (1995) manipulated the variance; however, they did

not investigate TIE. Instead, they compared the true population Mantel-Haenszel common odds ratio conditioned on observed score (α_{MH_X}) versus that conditioned on latent ability (α_{MH_θ}). They found no difference. They investigated a concept that the present study explores: a single index of congruence in the ability distributions may provide a meaningful index of the combined effect due to unequal means and unequal variances. For each of the six levels of FG distribution [$N(0,1)$, $N(0,.5)$, $N(-1.5,1)$, $N(-1.5,.5)$, $N(-3,1)$, and $N(-3,.5)$], they mapped the degree of overlap between the RG and FG ability distributions to a scalar by defining the proportion of overlap as:

$$\int_{-\infty}^{\infty} \text{MIN}[g_R(\theta), g_F(\theta)]d\theta \, .$$

The present study examines the effect of unequal variances in the ability distributions on the TIE rate of the χ^2_{MH} test for detecting DIF. The effects of unequal means in the ability distributions, test length, total sample size, ratio of RG to FG sample sizes, studied item discrimination, studied item difficulty, and the interactions among these factors, are also investigated. Furthermore, this study attempts to systematically disentangle the effect of total sample size from that of sample size ratio; something not accomplished in the existing literature. Narayanan and Swaminathan (1994) noted that sample size ratio, as well total sample size, may have an effect on the power of the χ^2_{MH} test for detecting DIF, since detection increased more in their study when the small FG size was increased than when the large RG size was increased. However, in their study only two of the nine studied RG/FG sample size combinations had the same total sample size (300/300, and 500/100); therefore, the effects of total sample size and sample size ratio were not separated. This study explores the relationship between the TIE of the χ^2_{MH} test and the area of overlap in the ability distributions of the two groups. Sample size ratio

was of interest in this study because it was thought to play a role with the overlap of ability distributions in influencing the amount of overlap in observed scores.

Methods

Dichotomous item responses were simulated with the IRT three parameter logistic model (3PL), with a scaling constant (D) of 1.7. Core item responses were simulated by using estimated item parameters (Table 3) from the 3PL (D=1.7) calibration of the 20 math concepts items from the *ITBS*, third grade, 32-item mathematics concepts and estimation test. Calibration was performed with BILOG (Mislevy & Bock, 1990). Twenty core items, plus one studied item per test, produced a test length of 21 items representing a very short test. A medium-length test of 41 items was studied by duplicating the 20 core items. This method ensured that the characteristics of the core items, possibly associated with TIE inflation (e.g., average discrimination, Lu, 1996), were constant across levels of test length.

For all studied conditions, the ability distribution was simulated as $N(0,1)$ for the RG and $N(\mu_F,\sigma^2_F)$ for the FG, and 400 replications were performed. In all, 576 conditions were studied by fully crossing seven factors: three levels of studied item discrimination (a = 0.5, 1.0, 1.5), three levels of studied item difficulty (b = -1, 0, 1), two levels of equality of means in the ability distributions [equal (μ_F=0), and unequal (μ_F=-1)], four levels of RG/FG variance ratio in the ability distributions [1.0 (σ^2_F =1), 1.33 (σ^2_F =.75), 2.0 (σ^2_F =.50), and 4.0 (σ^2_F =.25)], two levels of test length (21 and 41 items), two levels of total sample size (1000 and 2000), and two levels

of RG/FG sample size ratio (1.0 and 3.0). The crossing of total sample size and sample size ratio produced four RG/FG sample size combinations (750/250, 500/500, 1500/500, and 1000/1000).

The crossing of the *a* and *b* parameters created nine studied items. The levels of the studied factors were chosen to reflect a realistic range occurring in actual tests, and were also ones frequently chosen in previous research (Table 1). For example, a recent 3PL ($D = 1.7$) calibration of multiple forms of the paper and pencil ACT mathematics test revealed that for a pool of 720 items, the mean (SD) of the *a* parameter was 1.02 (.33), and the mean (SD) of the *b* parameter was .16 (1.06) (Chang, 1998). In the present study, the *c* parameter was set to a constant value of .15 for the studied items, a reasonable value for either a five- or four-alternative multiple choice test, since the pseudo-guessing parameter is often somewhat lower than 1/(the number of alternatives). For example, the 20 math concepts core items estimated for this study had four alternatives, and the average estimated *c* value was .17. As for the studied values of the RG/FG ability variance ratio (VR), the value of 1.33 is approximately the mean observed score VR for the White/African-American comparison among primary- and secondary-level achievement scores (Table 2). Evidence suggests that 2.0 is a realistic upper bound for the RG/FG observed score VR in standardized achievement tests (Table 2). Nevertheless, larger ratios could arise in some contexts. For example, Hedges and Nowell (1995) reported two Male/Female VRs between 2.0 and 3.0 for vocational aptitude tests. Furthermore, Wilcox (1987) reported that of 14 articles he found in the *American Educational Research Journal* where a one-way ANOVA design was used, the VR exceeded 16.0 in three articles. However, he did not report the nature of the groups or the scales. Thus, it may be questioned whether those settings are ones in which DIF analyses would be performed. Therefore, the largest VR simulated in this

study was 4.0. The continuity correction in the χ^2_{MH} formula was employed. The two-stage purification process was not used.

The dependent variable was the dichotomous rejection status (0=retain H_o that no DIF exists, 1=reject H_o), from the χ^2_{MH} test, for each replication under simulated no-DIF conditions, at the .05 nominal level. Results were initially summarized statistically by using a full logit model to test all main and interaction effects (127 terms) at the .01 significance level. The SAS CATMOD Procedure was used for this (SAS Institute Inc., 1989). In order to examine statistically significant interactions, mean TIE proportions, averaged over levels of the non-interacting factors, were tabulated. Additionally, to point out which specific conditions under which combination of means and variances in the ability distributions displaved practically important TIE inflation, the TIE proportions were tabulated for all 576 conditions. Due to numerous conditions, the significance level for individual TIE proportions was chosen to be .001; therefore, the 99.9% normal-approximated confidence interval (CI) around the .05 nominal TIE proportion for a single condition was (.014, .086). For mean TIE proportions, the denominator used in the standard error formula for the 99.9% CI was equal to 400 multiplied by the number of conditions. TIE inflation was considered practically important if false rejection proportions exceeded .100, and practical deflation was said to occur when proportions were less than .01. Thus, the criterion for practical importance was more difficult to attain than statistical significance.

The proportion of overlap in the ability distributions was calculated by estimating the area of the intersection under the normally-distributed ability curves of the two groups. The area was calculated by summing miniature areas, which were estimated by multiplying a z-score

interval width of 0.1 by the height of the smaller curve at that z-score. The curves for the RG and FG are presented for the eight combinations of ability means and VRs (Figure 1).

Results

All main effects from the full logit model were statistically significant at the predetermined .01 significance level (Table 4). In fact, all main effects were significant beyond the .0001 level. The effect of RG/FG variance ratio in the ability distributions (χ^2=238.7) was greater than the effects of studied item difficulty (χ^2=86.2), sample size ratio (χ^2=63.2), and test length (χ^2=29.3), but less than the other three main effects. In particular, unequal means in the ability distributions had the largest effect (χ^2=602.1) followed by studied item discrimination (χ^2=434.2). The TIE proportions are displayed for levels of the main effects, averaged over levels of the other factors (Table 5). Consistent with previous research, TIE was greater on average for unequal means in the ability distributions (μ_F = -1), highly discriminating studied items (a = 1.5), large total sample size (2000), easy studied items (b = -1), and short test length (21 items). Unexamined in previous publications, TIE was greater on average for larger variance ratios (VRs) in the ability distributions and for a sample size ratio of 1.0. One might think that the larger RG/FG sample size ratio (3.0) would produce more inflation due to possibly greater mismatch in the observed score distribution. However, when total sample size was held constant, the 1.0 sample size ratio created greater TIE, probably due to the increased power from a larger FG. This result agrees with the comments noted above by Narayanan and Swaminathan (1994). The largest average TIE for the levels of the main effects, averaged over all other factors, did not exceed the pre-defined .10 level of practical importance.

However, the main effects need to be interpreted with caution since the logit model revealed several interactions (Table 4). Of the 29 statistically significant ($p<.01$) interactions, 21 were significant beyond the .001 significance level, and 15 were significant beyond the .0001 level. The 29 interactions could be explained by seven interactions that included all others (in bold type, Table 4). Consistent with previous literature, the interaction between studied item discrimination (*a*), studied item difficulty (*b*), unequal means, and total sample size (Table 6) revealed that when averaged over two levels of test length, two levels of sample size ratio and four levels of variance ratio, the mean TIE inflation was greatest and of practical importance for the highly discriminating easy item (a=1.5, b=-1), but only when the means of the ability distributions were unequal ($\mu_F = -1$), and more so for the total sample size of 2000 (average TIE=.215) than for 1000 (average TIE=.132). Furthermore, the mean TIE inflation for the lowly discriminating hard item (a=.5, b=1) was of mild practical importance, but only when the means were unequal and the total sample size was 2000 (average TIE=.106). However, unlike previous studies (Table 6), there was practically important average TIE inflation for the highly discriminating moderately difficult item (a=1.5, b=0), but only when the means were unequal and the total sample size was 2000 (average TIE=.141). Also unobserved previously, the average TIE inflation approached practical importance for the highly discriminating hard item (a=1.5, b=1), but only when the means were *equal* and the total sample size was 2000 (average TIE=.099). These last two results were due to the fact that mean TIE proportions were averaged over levels of VR. It will be shown that RG/FG VRs of 2 and above produce inflated TIE for these two items.

A similar pattern was observed for the interaction between the *a* parameter, the *b* parameter, unequal means, and test length (Table 7), substituting the effect of short test length

(21-items) for that of large total sample size (2000). However, the average TIE for the highly discriminating hard item (a=1.5, b=1), under *equal* means, did not approach practical importance for the 21-item test (.067). Therefore, the mean inflation for this item under *equal* means and unequal variances may be influenced more by large sample sizes than by short tests.

The interaction between studied item discrimination (a), studied item difficulty (b), equality of means, and VR was statistically strong (Table 4, #3, χ^2=80.3). Its examination (Table 8) was of central importance to this study, and will assist in explaining the interactions just discussed. One effect of *unequal means (μ_F=-1) and unequal variances (VR ≥ 1.33)* was to generate TIE inflation in the highly discriminating moderately difficult item (a=1.5, b=0) even though this item was not inflated in this or previous studies under unequal means alone (the exception was the study by Roussos and Stout (1996) but only when there were 3,000 examinees in each group). When averaged over total sample size, sample size ratio, and test length, the mean TIE inflation for this item was of practical importance when the variance ratio was 2.0 and the ability means were unequal (.125) . This inflation became worse when the VR was 4.0 (.194). The effect was also to exacerbate any TIE inflation that was already observed under unequal means alone for four items: the highly discriminating easy (a=1.5, b=-1) and lowly discriminating hard (a=.5, b=1) items mentioned above, as well as, but of less practical importance, for the moderately discriminating easy (a=1.0, b=-1) and lowly discriminating moderately difficult (a=.5, b=0) items. However, these patterns are best understood by examining the TIE for all 576 conditions, since the inflation was worse for the larger total sample size (2000), the short test (21-items), and the sample size ratio of 1.0 (Tables 13 and 14).

For example (Table 14B), under unequal means and unequal variances, as the VR increased from 1.0 to 4.0 for 1000 examinees in each group, the TIE increased gradually for the

lowly discriminating hard item (a=.5, b=1, TIE=.138 to .248) and the highly discriminating easy item (a=1.5, b=-1, TIE=.260 to .410), and only slightly for the moderately discriminating easy item (a=1.0, b=-1, TIE=.105 to .138). However, the TIE made a larger jump for the highly discriminating moderately difficult item (a=1.5, b=0, TIE=.090 to .413). The latter item yielded no practically important inflation under unequal means alone or under unequal variances alone for any sample size combination on either test length (Tables 13 and 14). In contrast, this item revealed practically important TIE inflation under the combination of unequal means and a VR of 2.0 for all four sample size combinations on the 21-item test (.113, .165, .158, and .248 for the sizes of 750/250, 500/500, 1500/500, and 1000/1000, respectively; Tables 13 and 14). The inflation for this item was practically important when the VR was 1.33 but only for 1000 examinees in each group (.145). The TIE for this item on the 41-item test for 1000 examinees in each group was not statistically inflated under unequal means alone (.068), but was practically inflated under the combination of unequal variances and unequal means (VR=2.0, TIE=.128; VR=4.0, TIE=.190) (Table 14B). On the 21-item test and the realistic sample sizes of 1500/500 (Table 14A), the TIE for this item was not statistically inflated (.068) under unequal means alone, and was around the nominal level when only the variances were unequal, even for a VR of 4.0 (TIE=.055), but was noticeably inflated under the combination of unequal means and unequal variances (VR=2.0, TIE=.158; VR=4.0, TIE=.260).

There were other instances where unequal variances in combination with unequal means combined to produce inflation, but of mild practical importance. For example (Table 14A), the easy item with medium discrimination (a=1.0, b=-1) on the 21-item test for sample sizes of 1500 and 500, showed no statistical TIE inflation (.083) under unequal means alone, but revealed practically important inflation (.103) under the combination of unequal means and a VR of 1.33.

The easy item with high discrimination (a=1.5, b=-1) on the 41-item test for 500 examinees in each group, revealed TIE that increased from statistical non-significance (.068) under unequal means alone to near practical importance (.098) when combined with a VR of 2.0, and to practical importance (.103) for a VR of 4.0 (Table 13B). The latter was the only occurrence of practically important TIE inflation on the 41-item test for sample sizes of 500 in each group. Additionally (Table 14A), the combination of unequal variances and unequal means for this item produced the only instance of practically important TIE inflation (.108) on the 41-item test when the sample sizes were 1500 and 500 and the VR was 2.0 (Table 14A).

The combination of *equal means (μ_F=0) and unequal variances (VR ≥ 1.33)* in the ability distributions had little effect on the average TIE for most items (Table 8). In fact, *equal ability means and a VR of 1.33* produced no practically important or statistical TIE inflation for any of the 576 conditions (Tables 13 and 14). However, two items showed practically important average TIE inflation under *equal ability means and a VR of 4.0,* when averaged over total sample size, sampling ratio, and test length (Tables 8): the highly discriminating easy item (a=1.5, b=-1, average TIE=.120) and the highly discriminating hard item (a=1.5, b=1, average TIE=.169). For single conditions (Tables 13 and 14), the highly discriminating easy item (a=1.5, b=-1) showed practically important inflation on three sample size combinations for the 21-item test and the 1000/1000 combination for the 41-item test, but only when the VR was 4.0 (TIE=.073, .128, .155, and .263 for sizes of 750/250, 500/500, 1500/500, and 1000/1000, respectively, on the 21-item test, and .053, .055, .088, and .150 for the same sizes on the 41-item test). The highly discriminating hard item (a=1.5, b=1), on the other hand, demonstrated TIE inflation of mild practical importance when the ability means were equal and the VR was 2.0, but only for the 21-item test (TIE=.075, .108, .108, and .113 for sizes of 750/250, 500/500,

1500/500, and 1000/1000, respectively). Moreover, when the ability means were equal and the VR was 4.0, this item yielded inflation of practical importance for all four sample size combinations on the 21-item test, and both sampling ratios for the 2000 total sample size on the 41-item test (TIE=.128, .178, .250, and .298 for sizes of 750/250, 500/500, 1500/500, and 1000/1000, respectively, on the 21-item test, and .088, .098, .133, and .183 on the 41-item test).

However, the most striking surprise concerning the TIE of average or single conditions was that the highly discriminating hard item (a=1.5, b=1) manifested the greatest TIE inflation under unequal variances alone, and yet showed no practically important or statistically significant inflation under unequal means alone, or even more interestingly under the combination of unequal means and unequal variances (Tables 8, 13, and 14), except for the inflation of mild practical importance (.103) under unequal ability means and a VR of 4.0 on the 21-item test with 1000 examinees in each group (Table 14B).

The interaction between VR and total sample size (Table 9) and between VR and test length (Table 10) revealed that the average TIE increased as VR increased, but the increase was greater for the total sample size of 2000 and the short test. Unexamined systematically in previous studies, the interaction between studied item discrimination (*a*), studied item difficulty (*b*), and sample size ratio (Table 11) showed that the sample size ratio of 1.0 produced greater average TIE than the 3.0 ratio, but the effect was larger for certain items: the lowly discriminating hard item (a=0.5, b=1), the highly discriminating easy item (a=1.5, b=-1), and the highly discriminating hard item (a=1.5, b=1). Additionally, the five-way interaction between sample size ratio, the *b* parameter, equality of means, total sample size, and test length was examined (Table 12). Averaged over VR and the *a* parameter, the sample size ratio of 1.0 produced greater practically important average TIE than the 3.0 ratio, but only under unequal

means in the ability distribution, and especially for the total sample size of 2000, and the 21-item test. The magnitude of average TIE under both sample size ratios was greatest for the easy items (b=-1).

Several other observations concerning Tables 13 and 14 are worth noting. None of the 576 conditions exhibited practically important or statistically significant (2-tailed p<.001) TIE *deflation* (TIE<.014). However, there were instances of statistical *mean* TIE *deflation* when averaged over the nine studied items, but only for the 21-item test and total sample size of 1000 under ability distributions that were *equal* in means with a VR of 1, 1.33, or 2. Ninety-one of the 576 conditions had a TIE that was statistically (2-tailed p<.001) *inflated* (>.086) above the .05 nominal level. Of these, 65 displayed an inflation that was practically important (TIE>.10). There were three items that showed no practically important TIE inflation for any of the 576 conditions: a=.5, b=-1; a=1.0, b=0, and a=1.0, b=1. The largest TIE proportions (.413 and .410) occurred for the easy item with high discrimination (a=1.5, b=-1) and the moderately difficult item with high discrimination (a=1.5, b=0), respectively, on the 21-item test with a total sample size of 2000, sample size ratio of 1.0, and ability distributions unequal in means (μ_F = -1) with a VR of 4.0 (Table 14B). Consistent with the literature, there was no practically important or statistical TIE inflation when the ability distributions were *equal in both means (μ_F = 0) and variances (σ^2_F = 1)*, for any of conditions (Tables 13 and 14).

The relationship between TIE and area of overlap (AO) in the ability distributions of the two groups was explored graphically for the total sample size of 2000. This was done separately for the four combinations of sample size ratio and test length (Figures 2-5). Only the six items that displayed at least one instance of practically important TIE inflation (>.10) are shown. The AO was 1.0, .93, .83 and .68 under VRs of 1.0, 1.33, 2.0, and 4.0, respectively, when the ability

distributions were equal in means. The AO for unequal ability means ($\mu_F = -1$) was .62, .59, .54, and .45 when combined with VRs of 1.0, 1.33, 2.0, and 4.0, respectively. Notice that the largest studied VR (4.0) under equal means yielded more overlap (AO=.68) than did a shift in means of one RG standard deviation under equal variances (AO=.62). The thick horizontal line in these figures demarcates the .10 level of practically important TIE inflation.

For 1000 examinees in each group and the 21-item test (Figure 2), the relationship between TIE and AO appeared to be monotonic decreasing for all studied items except for the highly discriminating hard item (a=1.5, b=1). The relationship was reasonably linear only for the highly discriminating easy item (a=1.5, b=-1). However, the relationship between AO and TIE appeared for all items to be remarkably linear within levels of the equality of means factor. However, the most pronounced example of non-linearity as the ability means shifted, was the anomalous behavior of the highly discriminating hard item (a=1.5, b=1). The TIE for this item peaked under *equal* means and a VR of 4.0, but then severely dropped when the means of the ability distributions became *unequal*, even though the AO decreased. The trends for these items were similar for the other three combinations of sample size ratio and test length for the total sample size of 2000 (Figures 3-5), even though the overall magnitude of TIE was less.

Conclusions

Because the Mantel-Haenszel chi-square (χ^2_{MH}) is one of the most popular tests for detecting DIF, it is important to know its characteristics such as Type I error and power, under a variety of conditions. If the χ^2_{MH} test has large Type I error under certain conditions, then practitioners should be warned about the potential for high false rejection rates under those

conditions. It was previously thought that Type I error rates were inflated under unequal mean ability distributions for short tests, but only for the occasional highly discriminating easy item and sometimes for the lowly discriminating hard item. The present study shows that if one considers the realistic condition of unequal variances in the ability distributions, then inflation may actually occur for a somewhat wider variety of items than if the ability distributions differed only in their means. There were several conditions for which unequal variances in the ability distributions, either alone or in combination with unequal means, produced practically important TIE inflation when no such inflation was observed under unequal means alone. This occurred primarily for the hard (b=1) and medium difficulty (b=0) items with high discrimination (a=1.5) and on the short test (21 items) with the total sample size of 2000, but was also observed for the medium-length test (41 items), primarily with the larger total sample size (2000). Moreover, when TIE inflation did occur under unequal means alone, as expected from previous research, for the easy item with high discrimination (a=1.5, b=-1), and for the hard item with low discrimination (a=.5, b=1), the inflation was worse when the ability distributions also differed on variances.

The most peculiar finding was the aberrant behavior of the highly discriminating hard item (a=1.5, b=1), which was the only item to show marked TIE inflation due to unequal variances alone when only rare mild inflation was observed under the combination of unequal variances and unequal means. All other items showed the greatest inflation under the combination of unequal means and unequal variances. Perhaps there is some interaction between item type and the amount of discrepancy in the RG and FG sample sizes within the strata of the observed score distribution. However, judging from the overlap in the ability distributions, it is not clear why there would be greater TIE inflation for a highly discriminating hard item when

μ_F=0 and VR=4 (Figure 1.D) than when μ_F=-1 and VR=4 (Figure 1.H). The sample size ratio of 1.0 produced greater TIE inflation than the 3.0 ratio, but primarily for the highly discriminating easy item (a=1.5, b=-1), the highly discriminating moderately difficult item (a=1.5, b=0), and the lowly discriminating hard item (a=.5, b=1), under unequal means in the ability distributions for the larger total sample size and the short test.

In studying the relationship between AO and TIE, it may be useful to simulate a larger VR (e.g., 8) and/or a smaller difference in mean ability (e.g., μ_F = -.5) such that the AO due to the VR under equal mean ability interleaves with the AO due to the VR under unequal mean ability. This study contains several other limitations. Item responses were generated using estimated parameters from an actual 20-item mathematics test. It is not clear whether the results would generalize to other tests. The present study examined only the TIE of the χ^2_{MH} test and not indices of DIF magnitude such as the estimated Mantel-Haenszel common odds ratio ($\hat{\alpha}_{MH}$) or its delta-metric log-transformation, the MH-D-DIF. Uttaro and Millsap (1994) pointed out the importance of analyzing both DIF magnitude and statistical tests. For example, in their study, for a 40 item test there were no statistically significant main or interaction predictors of TIE for the χ^2_{MH} test; however, there were significant interactions between the studied item discrimination (a) and the equality of means of the ability distributions, and between the pseudo-guessing parameter (c) and the equality of means, in predicting the magnitude of $\hat{\alpha}_{MH}$. One could also examine the χ^2_{MH} values and/or the ETS method for categorizing DIF (A, B, and C categories). Additionally, the present study investigated TIE at the .05 nominal level. Results may change somewhat if the .01 level were examined, since inflated TIE for the χ^2_{MH} test could be accompanied by an inflated variance as well as an inflated mean of the chi-square distribution, in which case greater misbehavior might be expected in the extremity of the right tail. Furthermore,

future studies ought to examine the effect of unequal variances on the power, as well as the TIE, of the χ^2_{MH} test for detecting DIF.

The pseudo-guessing parameter (*c*) was not studied. Previous studies (Lu, 1996; Roussos & Stout, 1996; Uttaro & Millsap, 1994) suggest that the TIE inflation in the present study would probably have been greater for the easy item with high discrimination if the pseudo-guessing parameter had been set to a value *less* than .15; and the TIE inflation would probably have been greater for the hard item with low discrimination if the pseudo-guessing parameter had been set to a value *greater* than .15.

One needs to be aware of the magnitude of shift in means when modeling unequal variances. For example, in the present study, *unequal* means in the ability distributions was modeled by simulating the RG ability as $N(0,1)$ and the FG ability as $N(-1, \sigma^2_F)$. Therefore, when the means were unequal and σ^2_F was .5, the mean FG ability was one RG standard deviation below the RG mean; however, this actually represented a shift in means of approximately 1.15 standard deviations based on the pooled variance when the sample sizes were equal. Therefore, if one desires a constant shift of one standard deviation based on the pooled variance, then for equal sample size conditions, one would, for example, simulate the FG ability as $N(-.866, .5)$ for the VR of 2.0, and $N(-.791, .25)$ for the VR of 4.0. Of course this would lead to larger corresponding areas of overlap. For example, the area of overlap for a VR of 2.0 would be .59 when FG~$N(-.866, .5)$ instead of .54 when FG~$N(-1, .5)$.

References

ACT. (1992). Highschool Profile Report . Iowa City: ACT.

Benbow, C. P. (1988). Sex differences in mathematical reasoning ability in intellectually talented preadolescents: their nature, effects, and possible causes. Behavioral and Brain Sciences, 11, 169-232.

Benbow, C. P., & Stanley, J. C. (1980). Sex differences in mathematical ability: fact or artifact? Science, 210, 1262-1264.

Benbow, C. P., & Stanley, J. C. (1983). Sex differences in mathematical reasoning ability: more facts. Science, 222, 1029-1031.

Bielinski, J., & Davison, M. (1998). Gender differences by item difficulty interactions in multiple-choice mathematics items. American Educational Research Journal, 35(3), 455-476.

Birch, M. W. (1964). The detection of partial association I: The 2 x 2 case. Journal of the Royal Statistical Society, B26, 313-324.

Birch, M. W. (1965). The detection of partial association II: The general case. Journal of the Royal Statistical Society, B27, 111-124.

Chang, S.-W. (1998). A comparative study of item exposure control methods in computerized adaptive testing. Unpublished Doctoral dissertation, The University of Iowa, Iowa City.

Donoghue, J. R., Holland, P. W., & Thayer, D. T. (1993). A Monte Carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential item functioning. In P. W. Holland & H. Wainer (Eds.), Differential Item Functioning (pp. 137-166). Hillsdale, NJ: Lawrence Erlbaum.

Feingold, A. (1992). Sex differences in variability in intellectual abilities: a new look at an old controversy. Review of Educational Research, 62, 61-84.

Feingold, A. (1994). Gender differences in variability in intellectual abilities: a cross-cultural perspective. Sex Roles, 30, 81-90.

Feingold, A. (1995). The additive effects of differences in central tendency and variability are important in comparisons between groups. American Psychologist, 50, 5-13.

Feldt, L. S., Forsyth, R. A., Ansley, T. N., & Alnot, S. D. (1993). Iowa Tests of Educational Development. Iowa City: The University of Iowa.

Hedges, L. V., & Friedman, L. (1993). Gender differences in variability in intellectual abilities: a reanalysis of Fiengold's results. Review of Educational Research, 63, 94-105.

Hedges, L. V., & Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. Science, 269, 41-45.

Holland, P. W. (1985, October). On the study of differential item performance without IRT. Paper presented at the Proceedings of the Military Testing Association.

Holland, P. W., & Thayer, D. T. (1985). An alternate definition of the ETS scale of item difficulty (Research Report RR-85-43). Princeton NJ: Educational Testing Service.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), Test Validity (pp. 129-145). Hillsdale NJ: Lawrence Erlbaum.

Hoover, H. D., Hieronymus, A. N., Frisbie, D. A., & Dunbar, S. B. (1993). Iowa Tests of Basic Skills. Chicago: The Riverside Publishing Company.

Humphreys, L. G. (1988). Sex differences in variability may be more important than sex differences in means. Behavioral and Brain Sciences, 11, 195-196.

Landis, J. R., Heyman, E. R., & Koch, G. G. (1978). Average partial association in three-way contingency tables: A review and discussion of alternative tests. International Statistical Review, 46(3), 237-254.

Lewis, C. (1993). A note on the value of including the studied item in the test score when analyzing test items for DIF. In P. W. Holland & H. Wainer (Eds.), Differential Item Functioning (pp. 317-319). Hillsdale, NJ: Lawrence Erlbaum.

Lord, F. M. (1977). A study of item bias using item characteristic curve theory. In Y. H. Poortinga (Ed.), Basic Problems in Cross-Cultural Psychology (pp. 19-29). Amsterdam: Swets & Zeitlinger.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum.

Lu, S.-M. (1996). The relationship between item statistics and the Mantel-Haenszel and standardization DIF statistics when comparison groups differ in ability. Unpublished Doctoral dissertation, The University of Iowa, Iowa City.

Mantel, N., & Byar, D. P. (1978). Marginal homogeneity, symmetry and independence. Communication in Statistics, A7, 953-976.

Mantel, N., & Haenszel, W. M. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22, 719-748.

Merideth, W., & Millsap, R. E. (1992). On the misuse of manifest variables in the detection of measurement bias. Psychometrika, 57, 289-311.

Mislevy, R. J., & Bock, R. D. (1990). BILOG: Item analysis and test scoring with binary logistic models. (2nd ed.). Mooresville, IN: Scientific Software.

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. Applied Psychological Measurement, 18(4), 315-328.

Pommerich, M., Spray, J. A., & Parshall, C. G. (1995). An analytical evaluation of two common-odds ratios as population indicators of DIF (ACT Research Report Series 95-1). Iowa City IA: ACT.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. Applied Psychological Measurement, 17(2), 105-116.

Roussos, L., & Stout, W. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. Journal of Educational Measurement, 33(2), 215-230.

SAS Institute Inc. (1989). SAS/STAT User's Guide, Version 6. (Fourth ed.). (Vol. 2). Cary, NC: SAS Institute Inc.

Shepard, L. A., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. Journal of Educational Statistics, 6, 317-375.

Shepard, L. A., Camilli, G., & Williams, D. M. (1984). Accounting for statistical artifacts in item bias research. Journal of Educational Statistics, 9, 93-128.

Uttaro, T., & Millsap, R. E. (1994). Factors influencing the Mantel-Haenszel procedure in the detection of differential item functioning. Applied Psychological Measurement, 18(1), 15-25.

Wilcox, R. R. (1987). New designs in analysis of variance. Annual Review of Psychology, 38, 29-60.

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? Journal of Educational Statistics, 15(3), 185-197.

Table 1. Characteristics of Previous Simulation Studies and the Present Study[1]

	Donoghue, Holland, & Thayer[2] (1993)	Rogers & Swaminathan (1993)	Narayanan & Swaminathan (1994)	Uttaro & Millsap (1994)	Roussos & Stout (1996) study 1	Roussos & Stout (1996) study 2	Lu (1996)	(present) Monahan (2000)
Simulation Model:	*modified* 3PL	2PL, 3PL	3PL	3PL	3PL	3PL	2PN	3PL
Dependent variable(s):	MH-D-DIF MH-P-DIF	χ^2_{MH} α_{MH}	TIE (.05), TIE(.01)	TIE (.05) α_{MH}	TIE (.05)	TIE (.05) ETS	TIE (.05), ETS, χ^2_{MH} STD-P-DIF, MH-D-DIF,	TIE (.05)
# of conditions	576	24	1,296	36	12	180	2,240	576
# of replications	100	100	100	200	400	100	100	400
RG/FG sample sizes	2000/500	250/250 500/500	RG(300,500,1000) crossed with FG(100,200,300)	500/500	100,200,500, and 1000 ea.	500,1000 and 3000 ea.	500/250 1000/500	1500/500, 750/250 1000/1000, 500/500
RG/FG sample size ratio	4	1	various	1	1	1	2	3, 1
Test length	40	40	40	20, 40	25	26	25, 50	21, 41
# or % of non-studied items with DIF	0, 1, 2, 4	0	10%, 20%	0	0	0	0	0
Purification[3]	No	No	Yes	No	No	No	No	No
Studied item,								
Discrimination (*a*)	.3, 1, 1.5	.6, 1, 1.6	.5, .9, 1.25	.5, 1, 1.5	1.32	.4, 1, 2.5	.2,.39,.48,.57,.64,.71,.84[4]	.5, 1, 1.5
Difficulty (*b*)	-.5, 0, .5	-1.5, 0, 1.5	not reported	0, .3, .5	0.03	-1.5,-.5,0,.5,1.5	.33,.49,.61,.74,.88[5]	-1, 0, 1
Pseudo-guessing (*c*)	.2	.2 for 3PL	.2	0, .2	.25	0.25	0, .2[6]	.15
Mean of FG ability	-1	1	0, -.5, -1	0, -1	μ_R-μ_F=0, .5, 1	μ_R-μ_F=0, 1	-1	0, -1
Variance of FG ability	1	1	1	1	1	1	1	1, .75, .5, .25
RG/FG variance ratio	1	1	1	1	1	1	1	1, 1.33, 2, 4

[1]Only studies relating to the Type I error of the χ^2_{MH}, or the magnitude of the α_{MH} under true no-DIF, are reported.

All studies simulated RG ability distribution as N(0,1), except Roussos and Stout chose μ_R and μ_F so midpoint between them = mean of *b*.

[2]Only the second part of the study where the studied item was included in the matching score is described.

[3]Two-stage process recommended by Holland and Thayer (1988): items showing DIF in the 1st run are "purified" from the matching score on the 2nd run.

[4]Biserial correlations. [5]Proportion correct. [6]Pseudo-guessing accounted for by converting 2PN-simulated responses of '0' to '1' with probability of c=0 or .2.

KEY: χ^2_{MH} = Mantel-Haenszel chi-square values. α_{MH} = estimated Mantel-Haenszel odds ratio. 3PL = IRT 3-parameter logistic model. 2PL = IRT 2-parameter logistic model. 2PN = IRT 2-parameter normal ogive model. TIE(.05) = Type I error at .05 alpha ETS = % of B or C items in ETS classification system. RG = Reference group. FG = Focal group.

Table 2. Ratio of Variances for Whites/African-Americans, and Males/Females, on Various Standardized Achievement Tests.

A. Primary Level.

		Whites/ African-Americans	Males/ Females
	Mean	1.37	1.11
	Standard Deviation	0.18	0.10
Iowa Tests of Basic Skills (ITBS) (Hoover et al., 1993)	Minimum	0.96	0.88
[n=178 combinations of grade (3-8) and test]	Percentiles: 5	1.06	0.96
	25	1.26	1.04
	50	1.38	1.11
	75	1.50	1.19
	95	1.65	1.26
	Maximum	1.87	1.40

B. Secondary Level.

	Mean	1.32	1.21
	Standard Deviation	0.16	0.09
Iowa Tests of Educational Development (ITED) (Feldt et al., 1993)	Minimum	0.91	0.99
	Percentiles: 5	1.07	1.05
[n=44 combinations of grade (9-12) and test]	25	1.23	1.15
	50	1.34	1.21
	75	1.42	1.28
	95	1.59	1.34
	Maximum	1.63	1.36

C. College Entrance.

	English	1.19	1.04
ACT (ACT, 1992)	Reading	1.35	1.10
(n=4 tests)	Science	1.78	1.25
	Math	1.80	1.23

Table 3. Parameters used to Simulate Item Responses for Non-studied (Core) Items, from the BILOG 3PL Calibration (D=1.7) of the *ITBS* Third Grade Mathematics Concepts Portion of the Mathematics Concepts and Estimation Test.

	Parameter		
Item No.	*a*	*b*	*c*
1	1.039	-2.382	0.145
2	0.752	-2.594	0.166
3	0.914	-1.229	0.144
4	1.058	-2.057	0.137
5	0.524	-1.345	0.140
6	0.910	-0.455	0.218
7	0.959	-2.177	0.135
8	1.027	-0.893	0.155
9	0.815	-0.066	0.273
10	0.505	0.989	0.230
11	0.749	-0.316	0.128
12	0.714	-0.935	0.150
13	0.571	0.608	0.235
14	0.584	-0.901	0.118
15	0.588	-0.069	0.298
16	0.597	-1.051	0.109
17	0.952	-1.274	0.140
18	1.070	0.267	0.177
19	1.610	-0.240	0.193
20	0.631	-1.101	0.172
Mean:	0.83	-0.86	0.17
SD:	0.27	0.98	0.05
Minimum:	0.51	-2.59	0.11
Q1:	0.59	-1.29	0.14
Q2:	0.78	-0.92	0.15
Q3:	0.98	-0.20	0.20
Maximum:	1.61	0.99	0.30

Table 4. Statistically Significant (p<.01) Main and Interaction Effects on Type I Error (from the full logit model consisting of 7 main effects and 120 interaction terms).

Abbreviation	Main Effects	Chi-square	df	p
(MEAN)	Equality of Means in the Ability Distributions	602.1	1	<.0001
(APAR)	Discrimination of Studied Item (a-parameter)	434.2	2	<.0001
(TOTSIZE)	Total Sample Size	293.1	1	<.0001
(VAR)	Reference/Focal group Variance Ratio in the Ability Distributions	238.7	3	<.0001
(BPAR)	Difficulty of Studied Item (b-parameter)	86.2	2	<.0001
(SAMRATIO)	Reference/Focal group Sample Size Ratio	63.2	1	<.0001
(LENGTH)	Test Length	29.3	1	<.0001

Examined in...	Interactions[1]			
	APAR X TOTSIZE	23.6	2	<.0001
	BPAR X TOTSIZE	20.8	2	<.0001
	BPAR X MEAN X TOTSIZE	13.3	2	0.001
Table 6	**APAR X BPAR X MEAN X TOTSIZE (#1)**	**17.1**	**4**	**0.002**
	APAR X LENGTH	48.3	2	<.0001
	MEAN X LENGTH	207.5	1	<.0001
	APAR X BPAR X LENGTH	43.5	4	0.0002
	APAR X MEAN X LENGTH	10.2	2	0.006
	BPAR X MEAN X LENGTH	15.5	2	0.0004
Table 7	**APAR X BPAR X MEAN X LENGTH (#2)**	**45.1**	**4**	**<.0001**
	APAR X VAR	175.8	6	<.0001
	APAR X MEAN	43.8	2	<.0001
	APAR X VAR X MEAN	30.7	6	<.0001
	APAR X BPAR	179.8	4	<.0001
	APAR X BPAR X MEAN	428.6	4	<.0001
	BPAR X MEAN	170.3	2	<.0001
	BPAR X VAR X MEAN	59.7	6	<.0001
Table 8	**APAR X BPAR X VAR X MEAN (#3)**	**80.3**	**12**	**<.0001**
Table 9	**VAR X TOTSIZE (#4)**	**19.7**	**3**	**0.0002**
Table 10	**VAR X LENGTH (#5)**	**48.3**	**2**	**<.0001**
Table 11	**APAR X BPAR X SAMRATIO (#6)**	**18.8**	**4**	**0.0009**
	MEAN X SAMRATIO	7.1	1	0.008
	SAMRATIO X TOTSIZE	19.1	1	<.0001
	MEAN X SAMRATIO X TOTSIZE	11.8	1	0.0006
	BPAR X MEAN X SAMRATIO X TOTSIZE	10.6	2	0.005
	BPAR X TOTSIZE X LENGTH	12.3	2	0.002
	SAMRATIO X TOTSIZE X LENGTH	14.6	1	0.0001
	MEAN X SAMRATIO X TOTSIZE X LENGTH	7.1	1	0.008
Table 12	**BPAR X MEAN X SAMRATIO X TOTSIZE X LENGTH (#7)**	**11.69**	**1**	**0.003**

[1]All of the 29 statistically significant interactions above are included within, and therefore explained by, the seven interactions in bold type.

Table 5. False Rejection (Type I Error) Proportions for Main Effects, Averaged over Levels of Other Factors.

Factor	Levels	No. of conditions averaged over	Mean Type I Error[2]
Means of the Ability Distributions[1]:	Equal (μ_F=0)	288	0.048
	Unequal (μ_F=-1)	288	**0.079**
Discrimination of Studied Item (a-parameter):	0.5	192	0.053
	1.0	192	0.049
	1.5	192	**0.088**
Total Sample Size:	1000	288	0.052
	2000	288	**0.075**
Reference/Focal group Variance Ratio in the Ability Distributions[1]:	1 (σ^2_F=1.0)	144	0.052
	1.33 (σ^2_F=.75)	144	**0.055**
	2 (σ^2_F=.50)	144	**0.064**
	4 (σ^2_F=.25)	144	**0.084**
Difficulty of Studied Item (b-parameter):	1	192	**0.059**
	0	192	**0.057**
	-1	192	**0.074**
Reference/Focal group Sample Size Ratio:	3	288	**0.057**
	1	288	**0.070**
Test Length:	41 Items	288	**0.055**
	21 Items	288	**0.072**

[1]The reference group ability distribution was simulated as $N(0,1)$ for all conditions. The focal group ability distribution was simulated as $N(\mu_F,\sigma^2_F)$.

[2]Mean Type I error proportions in bold type lie outside the 99.9% Confidence Interval (CI), and were therefore statistically different from the .05 nominal value at the 2-sided .001 significance level. The 99.9% normal-approximated CI was based on an estimated standard error equal to SQRT[(.05)(.95)/(400)(no. of conditions averaged over)]. The rounded CI was equal to (.048, .052) for the mean of 288 conditions, and (.047, .053) for the means of both 144 and 192 conditions.

Table 6. Examination of the Statistically Significant (p=.002) Interaction Between Studied Item Discrimination (*a*), Studied Item Difficulty (*b*), Equality of Means, and Total Sample Size on Type I Error at the .05 Nominal Level.

The Type I error for each cell is averaged over the 16 conditions of the remaining factors (two levels of test length, two levels of sample size ratio, and four levels of variance ratios).

		Total Sample Size			
		1000		2000	
Item type		Means of the Ability Distributions		Means of the Ability Distributions	
a	b	Equal (μ_E=0)	Unequal (μ_E=-1)	Equal (μ_E=0)	Unequal (μ_E=-1)
0.5	-1	**0.031**	0.048	0.051	0.057
0.5	0	**0.030**	0.058	0.041	**0.065**
0.5	1	**0.039**	**0.078**	**0.037**	0.106
1.0	-1	**0.034**	**0.060**	0.047	**0.092**
1.0	0	**0.030**	0.045	0.047	0.052
1.0	1	0.041	**0.039**	0.055	0.052
1.5	-1	0.043	0.132	**0.081**	0.215
1.5	0	**0.036**	**0.090**	0.053	0.141
1.5	1	**0.070**	**0.039**	**0.099**	0.057
	mean:	**0.039**	**0.065**	**0.057**	**0.093**

Note. Proportions below .041 (.047 for means) or above .059 (.053 for means) are statistically different from .05 at the 2-tailed .001 significance level (bold type).

Proportions above .100 are considered to represent practically important Type I error inflation in this study (shaded bold type).

Table 7. Examination of the Statistically Significant (p<.0001) Interaction Between Studied Item Discrimination (*a*), Studied Item Difficulty (*b*), Equality of Means, and Test Length on Type I Error at the .05 Nominal Level.

The Type I error for each cell is averaged over the 16 conditions of the remaining factors (two levels of total sample size, two levels of sample size ratio, and four levels of variance ratio).

		Test Length			
		21 Items		41 Items	
Item type		Means of the Ability Distributions		Means of the Ability Distributions	
a	b	Equal (μ_E=0)	Unequal (μ_E=-1)	Equal (μ_E=0)	Unequal (μ_E=-1)
0.5	-1	**0.037**	0.053	0.047	0.048
0.5	0	**0.028**	**0.070**	0.043	0.046
0.5	1	**0.035**	0.108	0.043	**0.063**
1.0	-1	**0.036**	**0.084**	0.045	**0.060**
1.0	0	**0.031**	0.046	0.043	0.042
1.0	1	**0.038**	0.048	0.048	0.041
1.5	-1	0.046	0.222	0.042	0.103
1.5	0	**0.040**	0.119	0.046	**0.068**
1.5	1	**0.067**	0.042	0.057	0.046
	mean:	**0.040**	**0.088**	**0.046**	**0.057**

Note. Proportions below .041 (.047 for means) or above .059 (.053 for means) are statistically different from .05 at the 2-tailed .001 significance level (bold type).

Proportions above .100 are considered to represent practically important Type I error inflation in this study (shaded bold type).

Table 8. Examination of the Statistically Significant (p<.0001) Interaction Between Studied Item Discrimination (*a*), Studied Item Difficulty (*b*), Equality of Means, and Equality of Variances on Type I Error at the .05 Nominal Level.

The Type I error for each cell is averaged over the eight conditions of the remaining factors (two levels of total sample size, two levels of sample size ratio, and two levels of test length).

Focal Group~:		*N*(0,1)	*N*(0,.75)	*N*(0,.5)	*N*(0,.25)	*N*(-1,1)	*N*(-1,.75)	*N*(-1,.5)	*N*(-1,.25)
Area Overlap (AO):		1.00	0.93	0.83	0.68	0.62	0.59	0.54	0.45
Item type		Equal Means (μ_F = 0)				Unequal Means (μ_F = -1)			
a	b	VR=1	VR=1.33	VR=2	VR=4	VR=1	VR=1.33	VR=2	VR=4
0.5	-1	0.040	0.040	0.044	0.040	0.048	0.052	0.053	0.057
0.5	0	0.038	**0.034**	**0.034**	**0.036**	0.055	0.061	0.062	**0.070**
0.5	1	0.037	0.038	0.041	**0.035**	**0.078**	**0.084**	**0.093**	0.112
1.0	-1	0.039	**0.034**	0.043	0.047	**0.066**	**0.074**	**0.078**	**0.086**
1.0	0	0.039	0.038	**0.036**	0.041	0.040	0.047	0.048	0.058
1.0	1	0.042	0.038	0.044	**0.067**	0.045	0.043	0.044	0.049
1.5	-1	0.039	0.039	0.049	0.120	0.146	0.157	0.178	0.213
1.5	0	0.039	0.041	0.047	0.052	0.062	**0.081**	0.125	0.194
1.5	1	0.038	0.045	**0.085**	0.169	0.043	0.044	0.045	0.059
	mean:	**0.039**	**0.039**	0.047	**0.067**	**0.065**	**0.071**	**0.081**	**0.100**

Note. VR=Reference/Focal group variance ratio in the ability distributions.

Proportions below .037 (.046 for means) or above .063 (.054 for means) are statistically different from .05 at the 2-tailed .001 significance level (bold type).

Proportions above .100 are considered to represent practically important Type I error inflation in this study (shaded bold type).

Table 9. Examination of the Statistically Significant (p=.0002) Interaction between Variance Ratio in the Ability Distributions and Total Sample Size on Type I Error at the .05 Nominal Level.

The Type I error for each cell is averaged over the 72 conditions of the remaining factors (two levels of test length, two levels of sample size ratio, two levels of equality of means, and nine studied items).

Reference/Focal Group	Total Sample Size	
Variance Ratio	1000	2000
1.00	**0.045**	**0.059**
1.33	0.047	**0.064**
2.00	0.053	**0.074**
4.00	**0.064**	**0.103**

Note. Proportions below .046 or above .054 are statistically different from .05 at the 2-tailed .001 significance level (bold type).

Proportions above .100 are considered to represent practically important Type I error inflation in this study (shaded bold type).

Table 10. Examination of the Statistically Significant (p<.0001) Interaction between Variance Ratio in the Ability Distributions and Test Length on Type I Error at the .05 Nominal Level.

The Type I error for each cell is averaged over the 72 conditions of the remaining factors (two levels of total sample size, two levels of sample size ratio, two levels of equality of means, and nine studied items).

Reference/Focal Group Variance Ratio	Test Length	
	21 Items	41 Items
1.00	**0.055**	0.049
1.33	**0.059**	0.051
2.00	**0.073**	0.054
4.00	**0.101**	**0.067**

Note. Proportions below .046 or above .054 are statistically different from .05 at the 2-tailed .001 significance level (bold type).

Proportions above .100 are considered to represent practically important Type I error inflation in this study (shaded bold type).

Table 11. Examination of the Statistically Significant (p=.0009) Interaction Between Studied Item Discrimination (*a*), Studied Item Difficulty (*b*), and Sample Size Ratio on Type I Error at the .05 Nominal Level.

The Type I error for each cell is averaged over the 32 conditions of the remaining factors
(two levels of total sample size, two levels of test length, two levels of equality of means, and four levels of variance ratio).

Item type		Sample Size Ratio	
a	b	1.0	3.0
0.5	-1	0.049	0.045
0.5	0	0.053	0.045
0.5	1	**0.076**	0.054
1.0	-1	**0.061**	0.055
1.0	0	0.047	**0.040**
1.0	1	0.051	**0.042**
1.5	-1	0.131	0.105
1.5	0	**0.093**	**0.068**
1.5	1	**0.069**	**0.063**
	mean:	**0.070**	**0.057**

Note. Proportions below .044 (.048 for means) or above .056 (.052 for means) are statistically different from .05 at the 2-tailed .001 significance level (bold type).

Proportions above .100 are considered to represent practically important Type I error inflation in this study (shaded bold type).

Table 12. Examination of the Statistically Significant (p=.003) Interaction between Studied Item Difficulty (*b*), Equality of Means, Sample Size Ratio, Total Sample Size, and Test Length on Type I Error at the .05 Nominal Level.

The Type I Error for each cell is averaged over the 12 conditions of the remaining factors (four levels of variance ratio, and three levels of studied item discrimination).

21 Items

	Total Sample Size							
	1000				2000			
	Sample Size Ratio				Sample Size Ratio			
	1.0		3.0		1.0		3.0	
	Means of the Ability Distributions		Means of the Ability Distributions		Means of the Ability Distributions		Means of the Ability Distributions	
b	Equal (μ_F=0)	Unequal (μ_F=-1)	Equal (μ_F=0)	Unequal (μ_F=-1)	Equal (μ_F=0)	Unequal (μ_F=-1)	Equal (μ_F=0)	Unequal (μ_F=-1)
-1	**0.034**	0.103	**0.031**	**0.093**	**0.070**	**0.170**	0.055	**0.141**
0	**0.032**	**0.086**	**0.024**	**0.071**	**0.037**	**0.129**	0.042	**0.089**
1	0.059	**0.062**	0.044	0.059	**0.066**	**0.103**	**0.066**	**0.063**

41 Items

	Total Sample Size							
	1000				2000			
	Sample Size Ratio				Sample Size Ratio			
	1.0		3.0		1.0		3.0	
	Means of the Ability Distributions		Means of the Ability Distributions		Means of the Ability Distributions		Means of the Ability Distributions	
b	Equal (μ_F=0)	Unequal (μ_F=-1)	Equal (μ_F=0)	Unequal (μ_F=-1)	Equal (μ_F=0)	Unequal (μ_F=-1)	Equal (μ_F=0)	Unequal (μ_F=-1)
-1	**0.038**	0.060	0.041	**0.065**	0.065	**0.102**	0.049	**0.072**
0	**0.034**	0.054	**0.038**	0.047	**0.064**	**0.075**	0.045	0.051
1	0.052	0.045	0.045	0.041	**0.065**	**0.071**	0.056	0.049

Note. Proportions below .040 or above .060 are statistically different from .05 at the 2-tailed .001 significance level (bold type). Proportions above .100 are considered to represent practically important Type I error inflation (shaded bold type).

Table 13. Empirical Type I Error at Alpha=.05 for *Total Sample Size of 1000 Examinees.*

A. Sample size ratio of 3.0 (750 Reference and 250 Focal Group Examinees)

		21-item test								41-item test							
Item type		Equal Means ($\mu_F = 0$)				Unequal Means ($\mu_F = -1$)				Equal Means ($\mu_F = 0$)				Unequal Means ($\mu_F = -1$)			
a	b	VR=1	VR=1.33	VR=2	VR=4	VR=1	VR=1.33	VR=2	VR=4	VR=1	VR=1.33	VR=2	VR=4	VR=1	VR=1.33	VR=2	VR=4
0.5	-1	.028	.018	.035	.023	.040	.048	.035	.033	.045	.043	.040	.033	.043	.058	.063	.065
0.5	0	.030	.023	.018	.015	.068	.068	.070	.078	.045	.043	.038	.040	.025	.040	.043	.048
0.5	1	.025	.030	.053	.033	.085	.078	.085	**.090**	.038	.038	.035	.033	.048	.053	.055	.055
1.0	-1	.023	.028	.028	.033	.048	.063	.073	.078	.058	.038	.038	.045	.043	.055	.060	.060
1.0	0	.018	.015	.020	.020	.043	.045	.043	.060	.038	.038	.038	.030	.038	.048	.048	.050
1.0	1	.020	.025	.035	.045	.038	.048	.040	.053	.040	.035	.043	.040	.038	.033	.033	.023
1.5	-1	.030	.025	.038	.073	.143	.163	.185	.208	.045	.030	.023	.053	**.088**	.078	.078	**.098**
1.5	0	.035	.028	.030	.033	.050	.073	.113	.140	.038	.033	.038	.040	.045	.048	.050	.080
1.5	1	.028	.038	.075	.128	.048	.040	.048	.053	.048	.050	.058	**.088**	.038	.043	.045	.035
mean		**.026**	**.025**	**.037**	.044	**.063**	**.069**	**.077**	**.088**	.044	.038	.039	.044	.045	.050	.053	.057

B. Sample size ratio of 1.0 (500 Reference and 500 Focal Group Examinees)

		21-item test								41-item test							
Item type		Equal Means ($\mu_F = 0$)				Unequal Means ($\mu_F = -1$)				Equal Means ($\mu_F = 0$)				Unequal Means ($\mu_F = -1$)			
a	b	VR=1	VR=1.33	VR=2	VR=4	VR=1	VR=1.33	VR=2	VR=4	VR=1	VR=1.33	VR=2	VR=4	VR=1	VR=1.33	VR=2	VR=4
0.5	-1	.023	.025	.025	.018	.053	.055	.053	.058	.038	.043	.033	.028	.045	.038	.038	.048
0.5	0	.025	.023	.025	.025	.065	.073	.068	.073	.040	.030	.030	.035	.045	.050	.058	.063
0.5	1	.038	.043	.040	.030	.083	**.098**	.105	.128	.053	.050	.048	.048	.063	.068	.075	.080
1.0	-1	.033	.020	.028	.023	.065	.063	.070	**.090**	.038	.033	.045	.043	.045	.045	.053	.055
1.0	0	.028	.025	.025	.040	.035	.038	.045	.058	.035	.038	.025	.045	.040	.045	.050	.033
1.0	1	.033	.028	.050	.073	.053	.043	.040	.053	.050	.045	.040	.050	.038	.030	.028	.033
1.5	-1	.030	.023	.033	.128	.160	.158	.188	.220	.035	.030	.038	.055	.068	**.088**	**.098**	.103
1.5	0	.035	.045	.050	.043	.063	**.093**	.165	.260	.038	.033	.020	.038	.053	.050	.068	**.098**
1.5	1	.033	.055	.108	.178	.030	.030	.033	.053	.035	.040	.068	**.098**	.033	.035	.028	.030
mean		**.031**	**.032**	.043	.062	**.067**	**.072**	**.085**	.110	.040	.038	.039	.049	.048	.050	.055	**.060**

Note. VR=Reference/Focal group variance ratio in the ability distributions.

Proportions below .014 (.038 for means) or above .086 (.062 for means) are statistically different from .05 at the 2-tailed .001 significance level (bold type).

Proportions above .100 are considered to represent practically important Type I error inflation in this study (shaded bold type).

Table 14. Empirical Type I Error at Alpha=.05 for *Total Sample Size of 2000 Examinees.*

A. Sample size ratio of 3.0 (1500 Reference and 500 Focal Group Examinees)

		21-item test								41-item test							
Item type		Equal Means ($\mu_F = 0$)				Unequal Means ($\mu_F = -1$)				Equal Means ($\mu_F = 0$)				Unequal Means ($\mu_F = -1$)			
a	b	VR=1	VR=1.33	VR=2	VR=4	VR=1	VR=1.33	VR=2	VR=4	VR=1	VR=1.33	VR=2	VR=4	VR=1	VR=1.33	VR=2	VR=4
0.5	-1	.038	.050	.050	.058	.053	.048	.063	.065	.050	.045	.050	.048	.045	.043	.043	.045
0.5	0	.040	.045	.033	.040	.058	.065	.065	.085	.038	.033	.038	.045	.045	.040	.045	.035
0.5	1	.033	.038	.038	.030	**.088**	**.093**	**.098**	.130	.030	.040	.040	.038	.033	.050	.055	.053
1.0	-1	.040	.038	.043	.048	.083	.103	.115	**.095**	.035	.033	.048	.065	.060	.070	.063	.068
1.0	0	.048	.035	.035	.045	.035	.053	.048	.080	.048	.043	.043	.040	.028	.040	.030	.038
1.0	1	.035	.040	.058	.085	.048	.040	.038	.035	.053	.045	.050	.063	.040	.048	.045	.048
1.5	-1	.040	.045	.060	.155	.230	.245	.273	.323	.043	.043	.038	**.088**	.085	**.090**	.108	.150
1.5	0	.040	.040	.050	.055	.068	**.098**	.158	.260	.030	.050	.065	.065	.063	.063	.075	.115
1.5	1	.038	.040	.108	.250	.045	.040	.043	.065	.055	.053	.078	.133	.055	.068	.045	.055
mean		.039	.041	.053	**.085**	**.079**	**.087**	**.100**	.126	.042	.043	.050	**.065**	.050	.057	.057	**.067**

B. Sample size ratio of 1.0 (1000 Reference and 1000 Focal Group Examinees)

		21-item test								41-item test							
Item type		Equal Means ($\mu_F = 0$)				Unequal Means ($\mu_F = -1$)				Equal Means ($\mu_F = 0$)				Unequal Means ($\mu_F = -1$)			
a	b	VR=1	VR=1.33	VR=2	VR=4	VR=1	VR=1.33	VR=2	VR=4	VR=1	VR=1.33	VR=2	VR=4	VR=1	VR=1.33	VR=2	VR=4
0.5	-1	.048	.040	.050	.050	.058	.068	.073	.080	.050	.058	.068	.065	.053	.060	.055	.065
0.5	0	.025	.025	.033	.033	.078	.083	**.093**	.123	.060	.055	.060	.055	.055	.070	.055	.055
0.5	1	.025	.020	.030	.023	.138	.148	.185	.248	.055	.050	.048	.050	.085	**.090**	**.088**	.115
1.0	-1	.043	.038	.055	.060	.105	.113	.110	.138	.045	.050	.058	.063	.083	.080	.073	.105
1.0	0	.040	.045	.035	.043	.050	.060	.070	**.095**	.058	.065	.065	.065	.053	.050	.050	.050
1.0	1	.043	.038	.033	**.098**	.063	.060	.070	**.090**	.060	.050	.045	.085	.045	.045	.060	.063
1.5	-1	.043	.060	.095	.263	.260	.288	.335	.410	.048	.060	.068	**.150**	.138	.150	.165	.198
1.5	0	.038	.035	.040	.058	**.090**	.145	.248	.413	.060	.068	.080	.083	.068	.083	.128	.190
1.5	1	.035	.045	.113	.298	.038	.040	.055	.103	.038	.043	.078	.183	.060	.060	.065	.080
mean		.038	.038	.054	.103	**.098**	**.111**	**.139**	**.189**	.053	.055	**.063**	**.089**	**.071**	**.076**	**.082**	**.102**

Note. VR=Reference/Focal group variance ratio in the ability distributions.

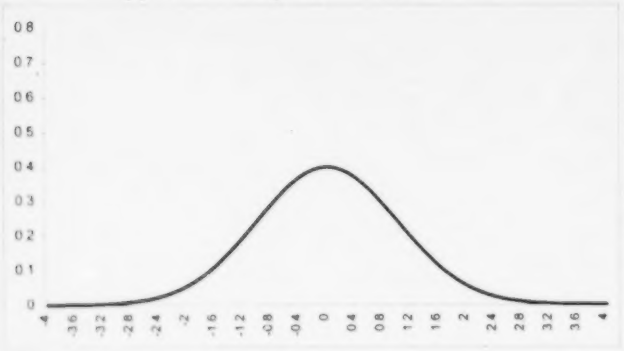
Proportions below .014 (.038 for means) or above .086 (.062 for means) are statistically different from .05 at the 2-tailed .001 significance level (bold type).
Proportions above .100 are considered to represent practically important Type I error inflation in this study (shaded bold type).

Figure 1. Overlap in the Ability Distributions Under Normal Curves.

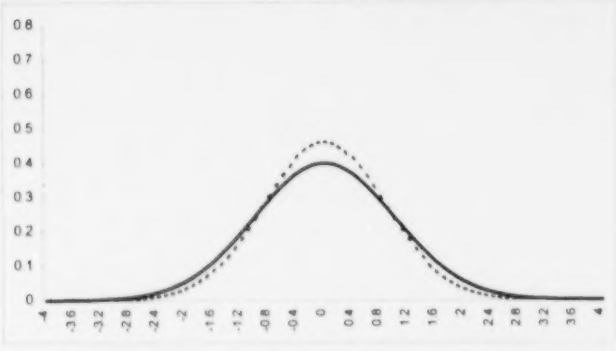
Solid line: Reference Group ~ N(0,1) Y-axis: Height under normal curve

Dashed line: Focal Group ~ N(μ_F,s^2_F) X-axis: Ability VR: Reference/Focal Group Variance Ratio

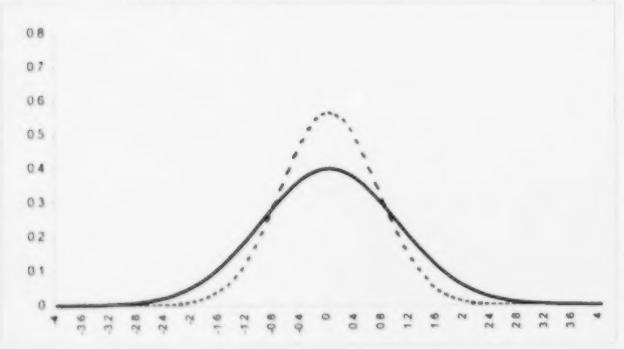
A. $\mu_F = 0$ $\sigma^2_F = 1$ VR = 1



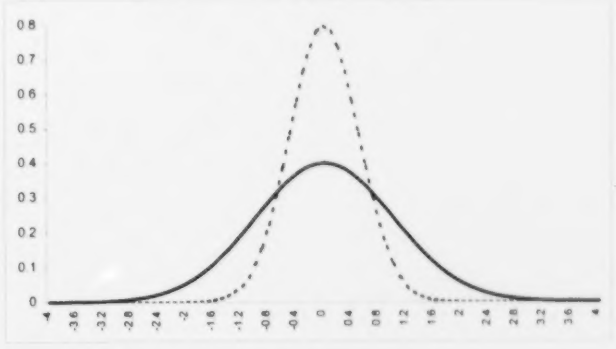
B. $\mu_F = 0$ $\sigma^2_F = .75$ VR = 1.33



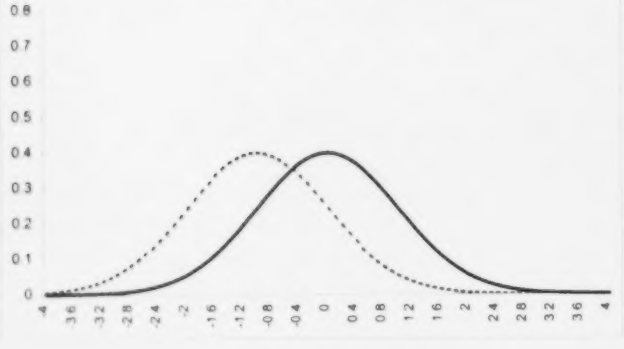
C. $\mu_F = 0$ $\sigma^2_F = .5$ VR = 2



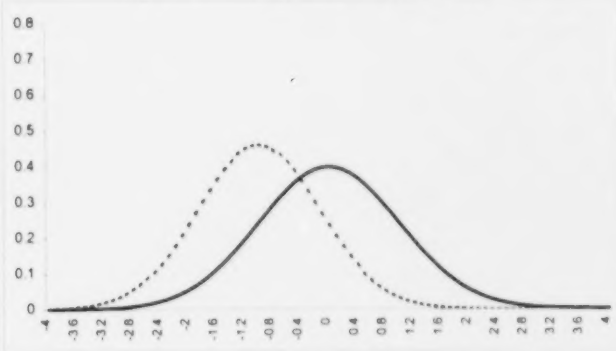
D. $\mu_F = 0$ $\sigma^2_F = .25$ VR = 4



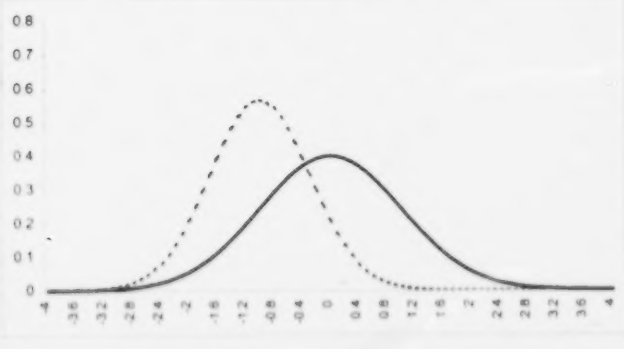
E. $\mu_F = -1$ $\sigma^2_F = 1$ VR = 1



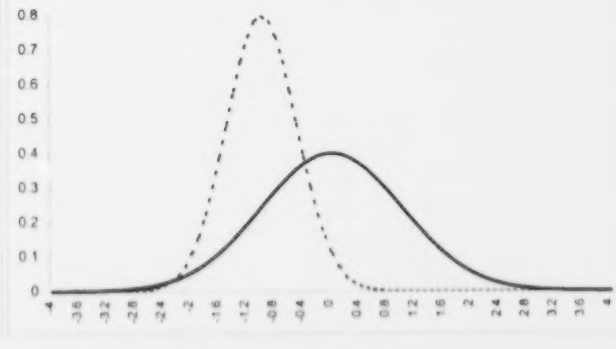
F. $\mu_F = -1$ $\sigma^2_F = .75$ VR = 1.33



G. $\mu_F = -1$ $\sigma^2_F = .5$ VR = 2



H. $\mu_F = -1$ $\sigma^2_F = .25$ VR = 4



BEST COPY AVAILABLE

Figure 2. Type I Error by Area of Overlap (AO) in Ability Distributions for 1000 Examinees in each Group and a 21-Item Test

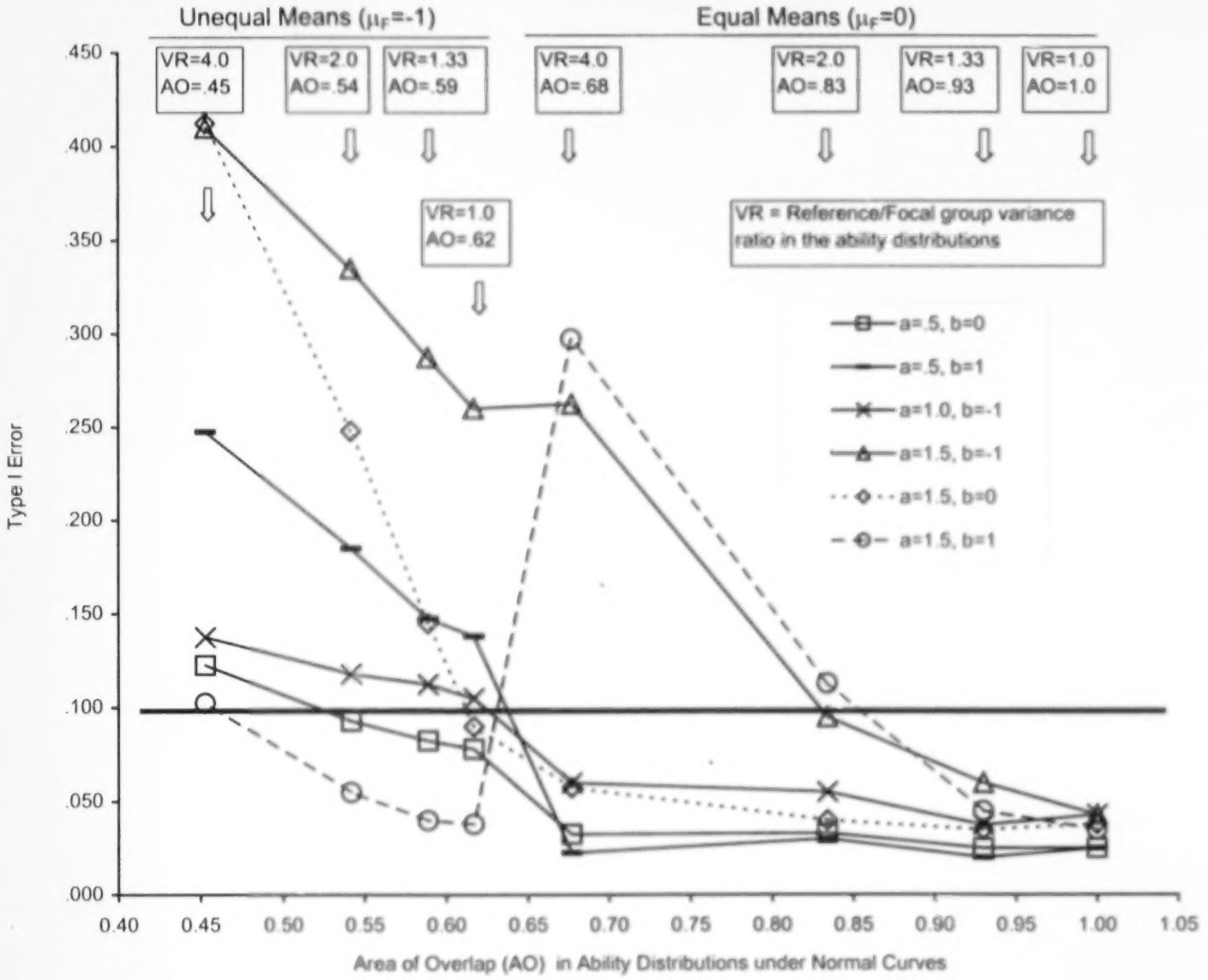


Figure 3. Type I Error by Area of Overlap (AO) in Ability Distributions for 1500 Reference and 500 Focal Group Examinees and a 21-Item Test

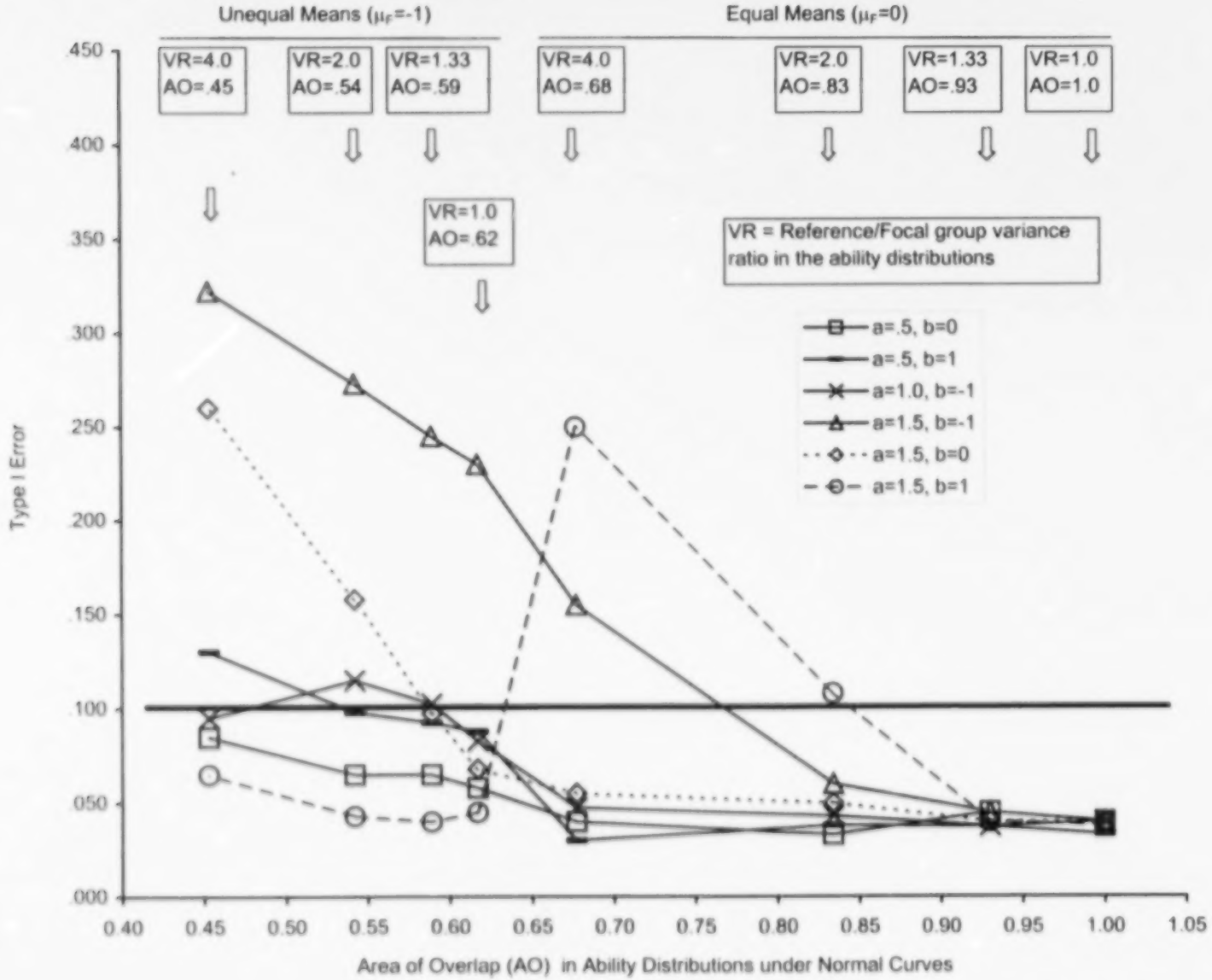


Figure 4. Type I Error by Area of Overlap (AO) in Ability Distributions for 1000 Examinees in each Group and a 41-Item Test

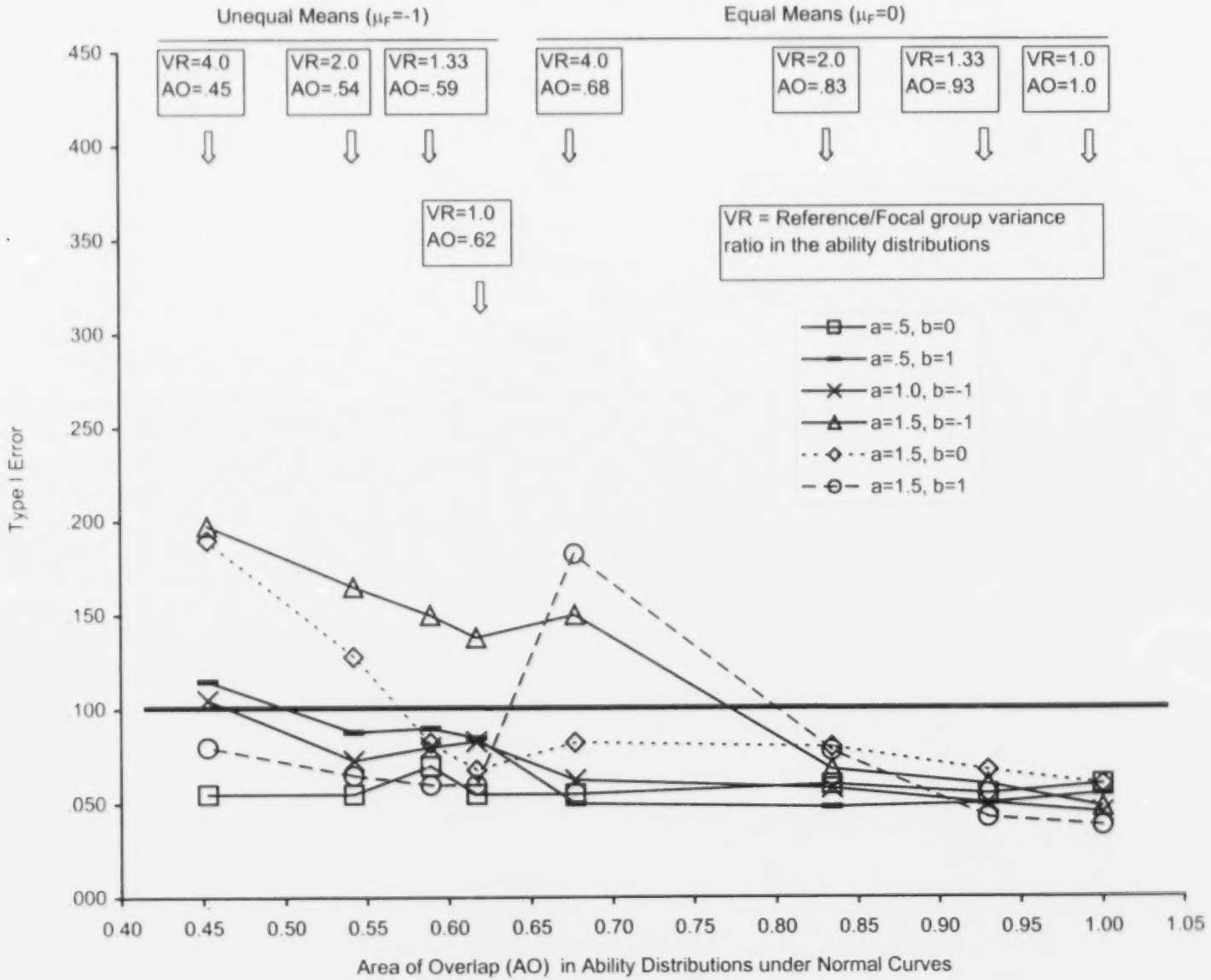


Figure 5. Type I Error by Area of Overlap (AO) in Ability Distributions for 1500 Reference and 500 Focal Group Examinees and a 41-Item Test